Artificial Intelligence in Medicine 79 (2017) 1–8

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

journal homepage: www.elsevier.com/locate/aiim

Protein fold recognition based on sparse representation based classification

Ke Yan[a], Yong Xu[a,*], Xiaozhao Fang[a], Chunhou Zheng[b], Bin Liu[a,*]

[a] School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, Guangdong, 518055, China
[b] College of Electrical Engineering and Automation, Anhui University, Hefei, Anhui, 230039, China

ARTICLE INFO

Article history:
Received 31 December 2016
Received in revised form 6 March 2017
Accepted 7 March 2017

Keywords:
Protein fold recognition
Sparse representation based classification
Protein representation

ABSTRACT

Knowledge of protein fold type is critical for determining the protein structure and function. Because of its importance, several computational methods for fold recognition have been proposed. Most of them are based on well-known machine learning techniques, such as Support Vector Machines (SVMs), Artificial Neural Network (ANN), etc. Although these machine learning methods play a role in stimulating the development of this important area, new techniques are still needed to further improve the predictive performance for fold recognition. Sparse Representation based Classification (SRC) has been widely used in image processing, and shows better performance than other related machine learning methods. In this study, we apply the SRC to solve the protein fold recognition problem. Experimental results on a widely used benchmark dataset show that the proposed method is able to improve the performance of some basic classifiers and three state-of-the-art methods to feature selection, including autocross-covariance (ACC) fold, D-D, and Bi-gram. Finally, we propose a novel computational predictor called MF-SRC for fold recognition by combining these three features into the framework of SRC to achieve further performance improvement. Compared with other computational methods in this field on DD dataset, EDD dataset and TG dataset, the proposed method achieves stable performance by reducing the influence of the noise in the dataset. It is anticipated that the proposed predictor may become a useful high throughput tool for large-scale fold recognition or at least, play a complementary role to the existing predictors in this regard.

1. Introduction

Protein fold recognition is crucial for predicting protein structure and function, which is one of the most important tasks in bioinformatics [1,2]. Fold recognition refers to recognition of structural fold of a protein based on the given sequence information, which is important for protein tertiary structure identification [3]. Most of the computational methods are based on machine learning techniques for fold recognition. There are two important components in these methods, feature extraction and classification. In this regard, several computational predictors have been proposed considering both the two important components.

During the past decades, many powerful feature extraction methods have been proposed. The early methods are based on the primary sequence of amino acids [4]. Some traditional methods utilize the amino acid composition features, such as n-gram composition, dipeptide composition, etc [5,6]. Taguchi and Gromiha [7] propose the syntactical based features such as occurrence and composition to represent the proteins. However, researchers have found that proteins sharing similar structures may only have low sequence similarities. Therefore, these sequence-based methods cannot perform well when the sequence similarity is low. To overcome this disadvantage, several methods incorporate the evolutionary information or structural information into the feature extraction process. Dubchak et al., [8] propose a new feature vector based on the physicochemical properties and structural of amino acids describing the structures, which is associated with the local and global information about amino acid sequence. This method is further improved by some other related studies [9–12]. In order to incorporate the sequence-order information into the predictor, Shen et al., [13] propose a computational method called ensemble classifier, which is based on the pseudo-amino acid composition (PseAAC), physicochemical features, predicted secondary structure of protein. Some popular methods extract the evolution information by using the PSI-BLAST [14] tool and show a better performance on the protein fold recognition. Dong et al., [15]

[*] Corresponding authors.
E-mail addresses: yanke401@163.com (K. Yan), laterfall@hitsz.edu.cn (Y. Xu), xzhfang168@126.com (X. Fang), zhengch99@126.com (C. Zheng), bliu@insun.hit.edu.cn (B. Liu).

http://dx.doi.org/10.1016/j.artmed.2017.03.006
0933-3657/© 2017 Elsevier B.V. All rights reserved.

combine the autocross covariance and PSSM to transform the protein sequences into vectors with fixed-length. Recently, a Hidden Markov Model (HMM) combines the Multiple Sequence Alignment (MSA) to incorporate the evolution information. Remmert et al., [16] propose an effective tool HHblits to perform the remote protein detection. Lyons et al. combine the HHblits and dynamic programming to perform the protein fold recognition [17]. These methods are only based on protein sequence composition information, the physicochemical properties, and evolutionary information. Zakeri et al., [18] propose the functional information, which is effective to improve the performance. But functional domain information is usually extracted by experimental methods or by known structural information [17]. Most of the aforementioned features are complementary. Therefore, several methods combine multiple features into a predictors, and performance improvement can be observed [18–20]. For more information, please refer to a recent review paper on fold recognition [21].

Another important component of computational predictors for fold recognition is the classification algorithm. Some well-known machine learning techniques have been applied to this field, such as Support Vector Machine (SVM) [15,22–32], linear discriminant analysis (LDA) [33], the artificial neural network (ANN) [34–36], k-nearest neighbor (KNN) [37], Bayesian network [38], random forest [39–42], etc. These methods treat fold recognition as a multi-class classification task. Among these methods, the SVM-based method can achieve the state-of-the-art performance. SVM has been successfully applied to the classification and regression tasks, which calculates the maximum margin hyperplane among the training samples to minimize classification error. The kernel function is used to project the data from the original space into a new feature space. The SVM's performance depends on the kernel function, which quantifies the similarity between the protein sequences. The speed of convergence of SVM is faster than some methods, such as ANN [22]. The kernel function is connected with the discriminative features and the prior knowledge of the source data [43]. There are many kernel functions, such as gaussian kernel, polynomial kernel, radial basis function, etc. It is essential to select a suitable kernel in SVM. However, it is difficult to find a suitable kernel function in the applications. Recently, Zakeri et al., [18] combine the geometry means and different kernel matrices to improve the performance of the SVM-based method. Hu et al., [45] combine multi-view feature sets and ensemble classifier to solve the protein crystallization prediction problem.

Sparse Representation based Classification (SRC) [44,46] is a robust machine learning technique, which is stabile for feature selection classification tasks, and outperforms some traditional machine learning methods for some tasks in the field of image recognition and image processing, such as face recognition [47–49], texture classification [50], image denoising, image restoration, etc. Yu-An Huang et al., [51] propose a weighted sparse representation based classification (WSRC) method to solve the problem of protein-protein interactions (PPI). Dong-jun Yu et al. [52] combine the sparse representation technique with SVM, and improve the capability for predicting the binding residues. In these methods, a test sample is expressed by training samples of all classes via a linear representation. The coefficient matrix is sparse, and most nonzero elements in the matrix are essential for fold recognition. The substitution matrices obtained by training dataset and coefficients are used to predict the test sample directly. SRC uses the represent result to perform the final classification [44]. Motivated by its success, in this study, we apply the SRC to protein fold recognition. To improve the performance of the protein fold recognition, we combine some special features through the classifier SRC. Experimental results show that it can improve the predictive performance of some state-of-the-art methods.

2. Materials and methods

2.1. Dataset

Three datasets are used in the study to evaluate the performance of various computational predictors for fold recognition. Three datasets included DD dataset [13], EDD dataset [15] and TG dataset [53]. DD dataset contains 27folds which represent four major structure classes: α, β, α + β, and α/β. The training set has 311 sequences and the testing set contains 383 testing sequences whose sequence similarity is less than 35%. The sequences in the DD dataset were extracted from the Structural Classification of Protein (SCOP) version 1.63 [13].

The EDD dataset contains 3418 protein sequences which belong to the 27 different folds that essentially used in the DD dataset from SCOP (version 1.75), which has more sequences in the each fold [17]. The sequence identify between two proteins is no more than 40%. We use the EDD dataset to further evaluate our proposed method.

The third benchmark which is TG dataset, which contains 1612 protein sequences belonging to 30 different folds from SCOP (version 1.73) constructed by Taguchi and Gromiha [54]. The benchmark has the detailed information of the 30 different fold types is described in [53], and the sequence identify between two proteins is no more than 25%.

2.2. The processes of the competing methods

Three state-of-the-art methods, including ACC fold [15], Bi-gram [55], and D-D [22] are employed to validate whether the proposed SRC framework can improve their performance or not. All these methods are based on SVMs, and they employ different feature extraction methods. Among those methods, ACC fold and Bi-gram are profile-based methods, and the D-D is a sequence-based method. The detailed processes of these methods are shown in the followings.

2.2.1. ACC fold

ACC fold [15] applies the autocross-covariance transformation to extract the features from the PSSM. PSSM is a matrix with dimension of L*20, where L is primary sequence's length. Element $P_{i,j}\,(i \in [1, L], j \in [1, 20])$ of PSSM is interpreted as the probability of the j-th amino acid at the i-th position of protein sequence. The ACC fold transformation method is used to convert the PSSM matrix into a fixed length vector, with dimension of 400*LG (LG represents the distance between the amino acids in the PSSM) [56]. In this study, the value of LG is set as 4.

The process of ACC fold method is as follows. Firstly, the protein sequences' PSSM entries are calculated by the PSI-BLAST tool, which is directly associated with the evolutionary information. Secondly, the corresponding ACC matrix is obtained by the PSSM. The ACC matrix contains the two components: the AC (between the same property) and CC (between two different properties). AC is applied to measure the correlation of two same properties, which have the distances of LG along the sequence, and CC measures the correlation of two different properties between the distances of LG along the sequence [15]. The value of ACC is calculated by Eq. (1) and Eq. (2). Finally, the resulting feature vectors ACC are fed into SVM for classification.

$$AC\,(i, LG) = \Sigma_{j=1}^{L-LG}\left(P_{i,j} - \overline{P_i}\right)\left(P_{i,j+LG}\overline{P_i}\right) / \left(L - LG\right) \tag{1}$$

$$CC\,(i_1, i_2, \mathrm{LG}) = \Sigma_{j=1}^{L-LG}\left(P_{i_1,j} - \overline{P_{i_1}}\right)\left(P_{i_2,j+LG}\overline{P_{i_2}}\right) / \left(L - LG\right) \tag{2}$$

where $\overline{P_i} = \Sigma_{j=1}^{L} P_{i,j}/L,,\ \overline{P_i}$ is the average score of an amino acid i in the total protein sequence [15].

Table 1
The description and dimension of the amino acids composition and physicochemical features.

Feature name	Dimension
Amino acid composition	20
Hydrophobicity	21
Normalized Van der Waals volume	21
Polarity	21
Secondary structure	21
Polarizability	21

Table 2
Accuracies (%) for various features and classifiers on DD Dataset.

Methods	Q	Reference
D-D	56.5%	[22]
Bi-grams	69.5%	[55]
ACC fold	70.1%	[15]
SRC-DD	60.57%	This paper
SRC-Bigram	73.89%	This paper
SRC-ACC	71.20%	This paper

2.2.2. Bi-gram

Proteins often have the same subsequence as conserved regions when they are in the same folds. To evaluate the performance of different subsequences more reliably, Bi-gram [55] calculates the probability between the different amino acids in the adjacent rows in the PSSM matrix. The transform probability from amino acid mth to nth is measured by Eq. (3). The dimension of Bigram is 400. Finally, the predictor is constructed based on SVM. The bigram features are connected with the conserved regions of the protein directly [55].

$$Bigram_{m,n} = \sum_{i=1}^{L-1} P_{i,m}P_{i+1,n}, m, n\varepsilon [1, 20] \tag{3}$$

2.2.3. D-d

D-D method utilizes various sequences features to improve the protein fold recognition performance, including two types of structural models from the protein sequence, amino acid composition (C), protein secondary structure (S), and four types of physicochemical features, hydrophobicity (H), polarity (P), polarizability (Z), Normalized van der Waals volume (V). The name and the dimensions of different features are listed in Table 1 [22]. There six features have been widely used in the field of fold recognition.

2.3. Sparse representation based classification

Sparse representation based classification(SRC) [44] has received much attention in recent years. Especially, SRC is widely used in solving the difficult face recognition problems, such as occlusion, image corruption caused by the noisy pixels [44]. The problem of protein fold recognition is partially similar to face recognition, and the features of both these two tasks are high-dimensioned vectors.

SRC was first introduced by John Wright [44]. Sparse representation originated from the compressing sensing(CS) theory [57,58]. The training sample set is denoted as $D \in \mathbb{R}^{m\times n}$, and $y\varepsilon\mathbb{R}^{m}$ is a test sample. There are c classes, the training set D can be represented as

$$D = [D_1, D_2, \ldots, D_c] \tag{4}$$

where the m-th block D_m is defined as $D_m = \left[d_{m,1}, d_{m,2}, \ldots, d_{m,n_m}\right]$, $m\varepsilon[1, 2, \ldots, c]$, where n_m is the number of training setting of the m-th class. Test sample y with object m can approximately represented as a linear combination of the set D_m,

$$y = d_{m,1}x_{m,1} + d_{m,2}x_{m,2} + \ldots + d_{m,n_m}x_{m,n_m} \tag{5}$$

Eq. (5) also can be written by,

$$y = D_m x_m \tag{6}$$

where $x_m = \begin{pmatrix} x_{m,1}, & \cdots & x_{m,n_m} \end{pmatrix}^T$. It is an ideal case where the most entries of x are zeros except those entries corresponding to the m-th class are nonzero [44]. The ℓ_0 norm is a means to obtain a satisfactory solution. However, the corresponding issue is a NP-hard problem, and a unique solution cannot obtain since it is a nonconvex problem. The study [44] proposed that the ℓ_0 norm can be replaced by the ℓ_1 norm when the solution is sparse enough,

$$x = \underset{x\in\mathbb{R}^n}{argmin} \|Dx - b\|_2^2 + \lambda\|x\|_1 \tag{7}$$

As a result, the coefficients can be solved by Eq. (7). Given the solution from [44], the algorithm classifies the test sample depending on the following rules,

$$j = \underset{j}{argmin} ||y - D_j x_j||_2, j \in (1, 2, \ldots, c) \tag{8}$$

In summary, the SRC method represents a test sample as a sparse combination of training samples and then allocates it to the class corresponding to the minimum distance of all fold types' residuals between query sample and reconstructed sample of each fold type $D_j x_j$.

2.4. Performance measures

The accuracy is adopted to evaluate the performance of various methods, which is a common strategy for assessing the global performance of multi-class classification [15]. Q is defined as the ratio of the number of correctly predicted samples to all query samples

$$\mathrm{Q} = \frac{CN}{N} \times 100 \tag{9}$$

where CN is the number of protein samples whose fold types are correctly predicted and N is the total number of the test samples in the test dataset.

3. Results and discussion

3.1. Improving the performance of existing computational predictors via SRC

In order to validate whether the proposed SRC framework can improve the performance of the existing methods, three state-of-the-art methods are employed, including D-D [22], Bi-gram [55], and ACC fold [15]. All these three methods are based on the SVMs. In order to incorporate those methods into the framework of SRC, the SVMs are replaced by SRC. The experimental results of various methods on the DD benchmark dataset are listed in Table 2 and Fig. 1, from which we can see that the proposed SRC algorithm is able to improve the performance of all the three methods by 1–4% in terms of overall accuracy, indicating that the SRC is a suitable method for fold recognition.

We compare some others widely used machine learning classifiers with SRC on features ACC, Bigram, amino acids composition and physicochemical properties. We take experiments on DD-dataset, and the basic classifiers are KNN, Random Forest [39], libsvm [59], Naïve Bayes [60]. The detailed information of the predictive results is listed in Table 3. We can see from Table 3 that the best classifier is SRC in the three feature groups. The accuracy of SRC is improved by 1%–14%, indicating that SRC is an efficient approach for protein fold classification.

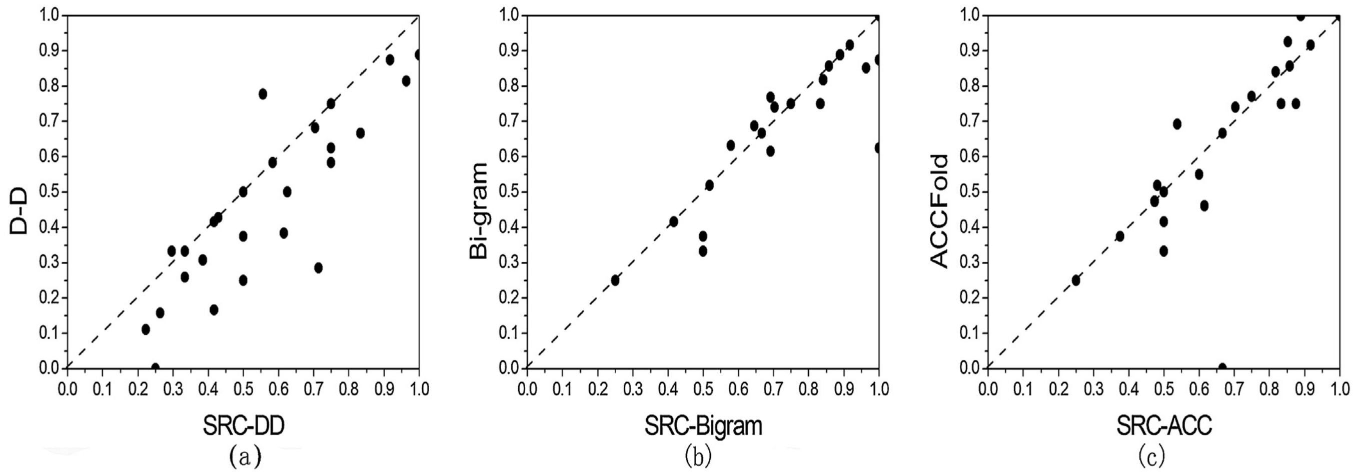

Fig. 1. Pairwise comparisons of SRC and SVM for DD benchmark dataset. The coordinates of each point in the plots are the accuracies of the SRC(x-axis) and SVM(y-axis) of each fold. The figures (a), (b), (C) show the comparisons using three features, including D-D, Bi-gram and ACC respectively. If a point falls in the up left part, it means that the method labeled in the y-axis outperform the method labeled in the x-axis. If a point falls into the diagonal, the performance of the two methods is identical for this fold.

Table 3
The Accuracy(%) of five basic classifiers on different features from DD-dataset.

Feature sets	Basic classifiers	Q
ACC	KNN	58.22%
	Random Forest	61.36%
	LibSVM	70.1%
	NaïveBayes	63.45%
	SRC(this paper)	71.2%
Bigram	KNN	59.01%
	Random Forest	63.19%
	LibSVM	69.5%
	NaïveBayes	56.4%
	SRC(this paper)	73.89%
Amino acids composition and physicochemical properties features	KNN	42.04%
	Random Forest	59.53%
	LibSVM	56.5%
	NaiveBayes	50.65%
	SRC(this paper)	60.57%

3.2. Combining various predictors into the framework of SRC

Combining complementary predictors is able to improve the performance. Therefore, it is interesting to explore whether the performance can be further improved by combining these three methods into the proposed SRC-based framework for fold recognition. The features generated by the three methods are linearly combined, and then are fed into the SRC for prediction. We call this new predictor as MF-SRC. The dimension of the corresponding feature vectors is $(20+21^{*}5+400+400^{*}LG)$, which can be represented as

$$F=\left[F_C, F_S, F_P, F_H, F_V, F_Z, F_{bi-gram}, F_{AAC}\right] \quad (10)$$

In order to help the readers to understand the process of MF-SRC, the flow chart is shown in Fig. 2. There are two main parts, including training phase and testing phase. In the training phase, sequence features are extracted as discussed in Section 2.2. A dictionary D (Eq. (4)) is constructed to store those features, and then they are fed into the SRC classifier for protein fold recognition. The sparse representation coefficients are obtained by the ℓ_1 norm minimization algorithm, and then we calculate the errors between the test sample and reconstructed samples of different fold types. The predicted result is the class corresponding to the minimum errors (Eq. (8)). The performance of MF-SRC is shown in Tables 4–6. Compared with the basic predictors and other famous classifiers, MF-SRC achieves the best performance.

Table 4
Performance comparison among different computational predictors on DD dataset.

Methods	Q	Reference
DD	56.5%	[22]
ACCFold_AC	68.6%	[15]
ACCFold_ACC	70.1%	[15]
Shamim	59.1%	[61]
LA	38.8%	[62]
SWPSSM	67.8%	[63]
PFP-Pred	62.1%	[13]
Tax-Fold	71.5%	[64]
PFP-FunDseqE	70.5%	[56]
Bi-grams	69.5%	[55]
HPFP	74.21%	[65]
PFPA	73.6%	[27]
ProFold	76.2%	[66]
MF-SRC	78.6%	This paper

Table 5
Performance comparison among different computational predictors on EDD dataset by 10-fold cross-validation.

Methods	Q	Reference
PF1	50.8%	[67]
PF2	49.9%	[67]
PF	53.4%	[1]
O	46.9%	[53]
DD	40.9%	[22]
ACC fold	85.9%	[17]
Mono-gram	76.9%	[68]
Bi gram	84.5%	[55]
HMMFold	86%	[54]
MF-SRC	86.2%	This paper

Table 6
Performance comparison among different computational predictors on TG dataset by 10-fold cross-validation.

Methods	Q	Reference
PF1	38.8%	[67]
PF2	38.8%	[67]
PF	43.1%	[1]
O	36.3%	[53]
DD	32%	[22]
ACC fold	66.4%	[17]
Mono-gram	58.8%	[68]
Bi gram	68.1%	[55]
Tri-gram	72.5%	[69]
Alignment method	74%	[70]
Paliwal et al	77%	[71]
MF-SRC	79.8%	This paper

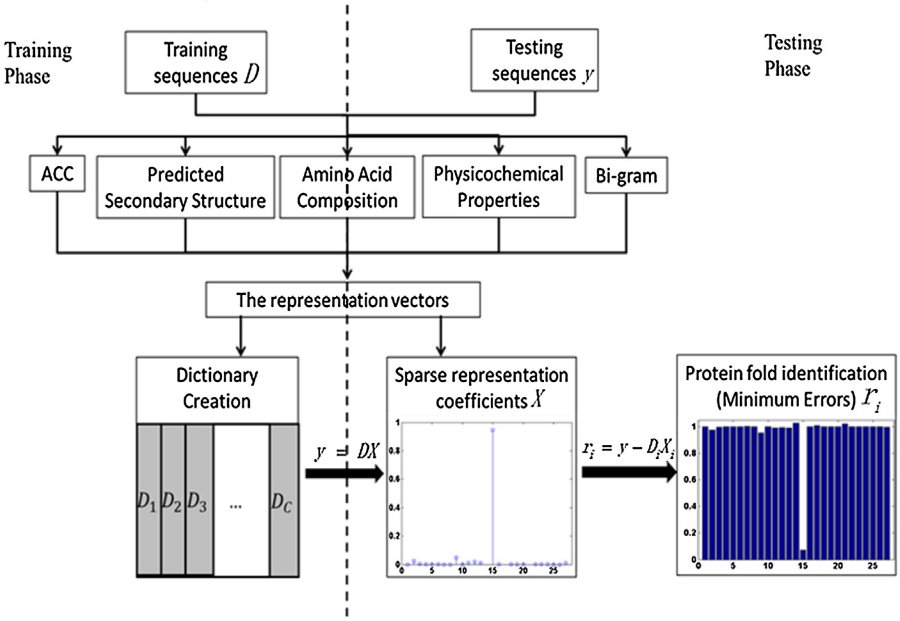

Fig. 2. The flowchart of the proposed MF-SRC framework. The primitive protein sequence can be converted into five features, including ACC, Bi-gram, Predicted second structure, Amino acid composition, physicochemical properties. Then the DD benchmark training dataset is converted into the features set D, and the test sequences are extracted by the same manner. A test sequence y is sparse represented by the training dataset D, and the sparse coefficients are illustrated in Fig. 2. Except that the entry from the fifteenth class, the most entries of coefficients are near zero. The residuals are calculated the distance between the test sample and constructed samples of each fold which are obtained by the training samples and coefficients by Eq. (7). Then the predicted label is assigned to the fifteen fold corresponding to the minimum residual.

3.3. Performance of MF-SRC

To further evaluate the effectiveness of MF-SRC, its performance is compared with some state-of-the-art methods. We first select the widely used DD dataset for evolution. Table 4 provides the total prediction accuracies of the existing methods for classification. As can be seen from Table 4, the accuracies of other methods are under 77%, and the MF-SRC accuracy is 78.6%, which is higher 2.4% than proFold [66]. The performance of MF-SRC outperforms all the other compared methods on the DD dataset.

Two other widely used benchmark datasets are used to further evaluate the performance of the proposed MF-SRC, including EDD and TG datasets. The performance of various methods on the EDD dataset and TG dataset is shown in Table 5 and Table 6, respectively. All these methods are evaluated by10-fold cross-validation. From these two tables we can see that the MF-SRC outperforms all the other compared methods. These experiments further demonstrate that the efficiency of the proposed MF-SRC.

3.4. Feature analysis of the MF-SRC

MF-SRC represents the target protein fold using a small number of atoms from the dictionary composed of all training samples. The sparse coefficient x is solved by Eq. (7). Usually most coefficients have values of zeros. To explore the reason why SRC works for protein fold recognition, a test sample in the fold type '4-helical cytokines' from the DD testing dataset is selected as an example. We obtain the sparse coefficients x by Eq. (7), and the results is shown in the Subfigure 3(a) . It is clearly that most values of coefficient are zero except for the coefficients from the fold type '4-helical cytokines', indicating that the most reconstructed samples through D_jx_j ($j\varepsilon$ [1, 2, . . ., c]) are different from the test sample apart from the fold type '4-helical cytokines'. We calculate the residuals to evaluate the distance between test sample and each reconstructed sample of 27 fold types by Eq. (8). Finally, the predicted result is the class with the minimum residual, which means the test sample can be represented by the training samples of this fold suitably. In other words, the test sample and the special training subset are the same fold. Subfigure 3(b) shows that the different residuals with respect to the 27 protein fold types and the predicted fold type of the test sample is '4-helical cytokines'.

SVM (or SVM based for ensemble strategy) classifiers have a quite promising results in protein fold recognition. The kernel function in the SVM can be regarded as the connection with the prior knowledge of source data and the discriminating features [43]. There are many kinds of kernels, such as tree kernel, polynomial kernel, Gaussian kernel, etc. Those kernel functions are designed for meeting different tasks. Selecting the best kernel among the existing candidates plays the most critical role in applying SVM. Unfortunately, it is not an easy task to find the suitable kernel functions. In contrast, MF-SRC utilizes the sparse representation of all training samples, instead of getting related models or features that can be used for classification [44]. As a result, the MF-SRC provides a succinct representation method for classification.

4. Conclusion

In the field of fold recognition, most of the discriminative methods are designed based on the well-known machine learning techniques, and most of these methods cannot perform well for fold recognition, because fold recognition is a multiple-classification task and the proteins share very low sequence similarities. Therefore, it is critical to explore advanced machine learning methods to further improve the predictive performance.

In this study, we apply the SRC to the field of fold recognition. In the step of feature selection, we extract the evaluationary features and physicochemical features. Experimental results show that the proposed method called MF-SRC outperforms some state-of-the-art methods. Evaluated on three widely used benchmark datasets, MF-SRC outperforms all the other compared methods, indicating that SRC is suitable for fold recognition. Compared with some state-of-the-art methods, we adopt a novel feature selection method and an effective predictor. As discussed in a recent review paper [13], homology detection is another important task in sequence

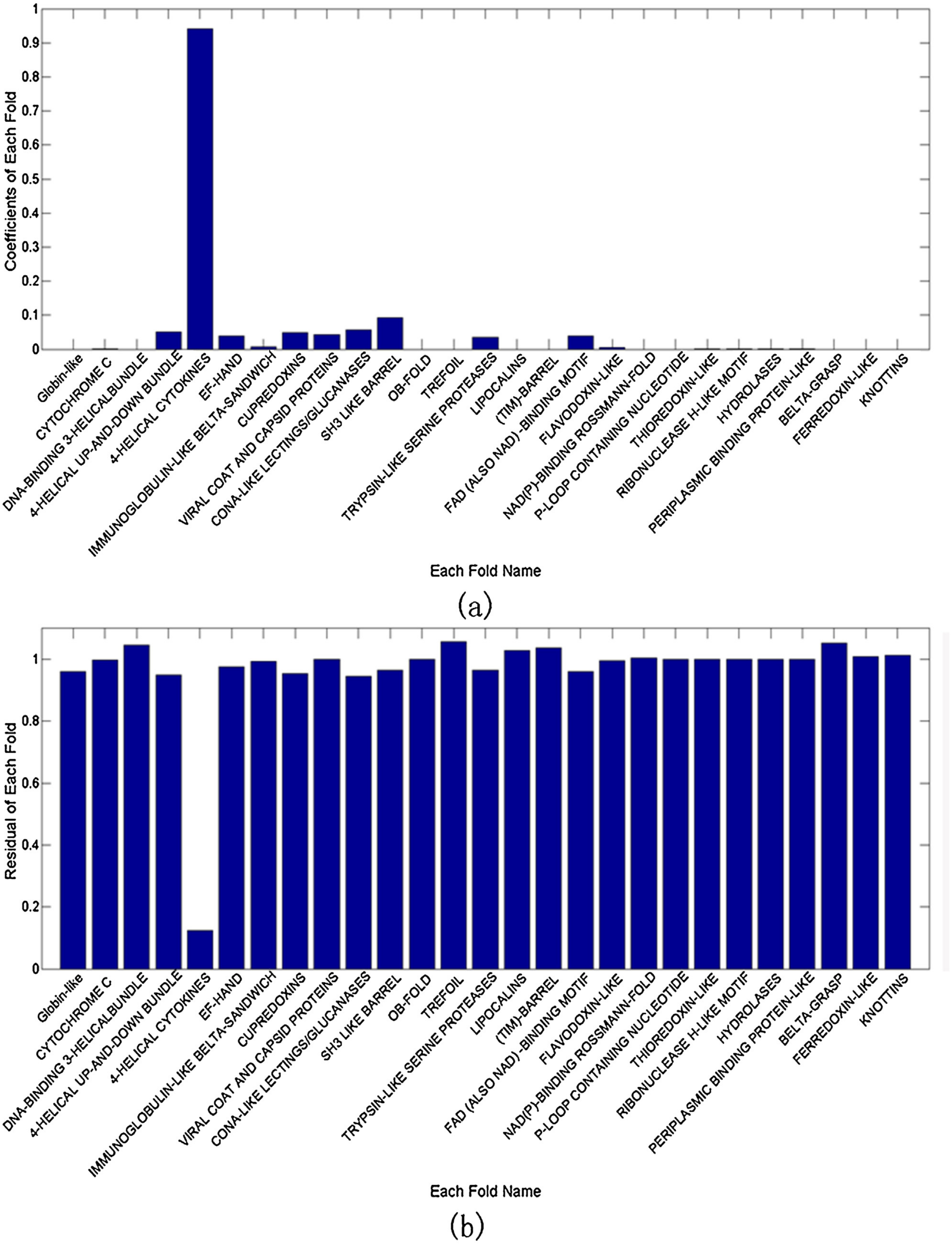

Fig. 3. An example to illustrate the reason why SRC works for fold recognition. A test sample in the fold type'4-helical cytokines' from the DD testing dataset is selected as an example. We calculate the squared sum of coefficients which are related to the 27 fold types, and the results are shown in Subfigure 3(a). The abscissa axis shows 27 fold types. The ordinate axis represents the sum of coefficient values of each fold type. To predict the fold type of test sample, we calculate the residuals of the 27 fold types, and the residual values of different fold types are shown in Subfigure 3(b). The x-axis shows 27 fold types, and the y-axis shows the residual values of different fold types. The fold type corresponding to the minimum residuals is the predicted result.

analysis. As shown in this study, SRC is able to improve the performance of protein fold recognition, and it can be used as a machine learning classification algorithm for binary and multiple tasks in bioinformatics, this task would be a potential application area of SRC. Future research will focus on exploring features with more discriminative power and advanced machine learning methods for fold recognition.

Conflict of interest statement

None declared.

Funding

This work was supported by the Shenzhen Municipal Science and Technology Innovation Council (Grant No. CXZZ20140904154910774 and Grant No. JCYJ 2014 0904154645958), National Natural Science Foundation of China (No. 61672184), Guangdong Natural Science Funds for Distinguished Young Scholars (2016A030306008) and Scientific Research Foundation in Shenzhen (Grant No. CYJ20150626110425228, and Grant No. JCYJ20150330155220591).

References

[1] Yang T, Kecman V, Cao L, Zhang C, Huang JZ. Margin-based ensemble classifier for protein fold recognition. Expert Syst Appl 2011;38:12348–55.

[2] M Craven, RJ Mural, LJ Hauser, EC Uberbacher, in ISMB. 98–106.

[3] J. Weston, C. Watkins, multi-class support vector machines (2016).

[4] Dubchak I, Muchnik I, Holbrook SR, Kim SH. Prediction of protein folding class using global description of amino acid sequence. Proc Natl Acad Sci 1995;92:8700–4.

[5] Dubchak I, Muchnik I, Holbrook SR, Kim S-H. Prediction of protein folding class using global description of amino acid sequence. Proc Natl Acad Sci 1995;92:8700–4.

[6] Nanni L, Lumini A, Gupta D, Garg A. Identifying bacterial virulent proteins by fusing a set of classifiers based on variants of Chou's pseudo amino acid composition and on evolutionary information. IEEE/ACM Trans Comput Biol Bioinf (TCBB) 2012;9:467–75.

[7] Taguchi YH, Gromiha MM. Application of amino acid occurrence for discriminating different folding types of globular proteins. Bmc Bioinf 2007;8:404.

[8] Dubchak I, Muchnik I, Kim S. Protein folding class predictor for SCOP: approach based on global descriptors. Proc Int Conf Intell Syst Mol Biol 1997;5:104–7.

[9] Pooya Z, Ben J, Raf V, Yves M. Protein fold recognition using geometric kernel data fusion. Bioinformatics 2014;30, 1850–1827.

[10] Y Krishnaraj, CK Reddy, in Bioinformatics and Biomedicine, 2008. BIBM '08. IEEE International Conference on. 393–96.

[11] Dehzangi A, Amnuaisuk SP, Ng KH, Mohandesi E. Protein fold prediction problem using ensemble of classifiers. Springer Berlin Heidelberg 2009.

[12] V Kecman, T Yang, in Computational Intelligence in Bioinformatics and Computational Biology, 2009. CIBCB '09. IEEE Symposium on. 75–8.

[13] Shen H-B, Chou K-C. Ensemble classifier for protein fold pattern recognition. Bioinformatics 2006;22:1717–22.

[14] Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST a new generation of protein database search programs. Nucleic Acids Res 25 1997:3389–402.

[15] Dong Q, Zhou S, Guan J. A new taxonomy-based protein fold recognition approach based on autocross-covariance transformation. Bioinformatics 2009;25:2655–62, http://dx.doi.org/10.1093/bioinformatics/btp500.

[16] Remmert M, Biegert A, Hauser A, Söding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods 2012;9:173–5.

[17] Lyons J, Paliwal KK, Dehzangi A, Heffernan R, Tsunoda T, Sharma A. Protein fold recognition using HMM?HMM alignment and dynamic programming. J Theor Biol 2016;393:67–74.

[18] Zakeri P, Jeuris B, Vandebril R. Protein fold recognition using geometric kernel data fusion. Bioinformatics 2014;30(13):1850–7.

[19] Ying Y, Huang K, Campbell C. Enhanced protein fold recognition through a novel data integration approach. Bmc Bioinf 2009;10:267.

[20] Chen J, Long R, Wang X, Liu B, Chou K-C. dRHP-PseRA: detecting remote homology proteins using profilebased pseudo protein sequence and rank aggregation. Sci Rep 2016;6:32333.

[21] Wei L, Zou Q. Recent progress in machine learning-Based methods for protein fold recognition. Int J Mol Sci 2016;17:2118.

[22] Ding CH, Dubchak I. Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics 2001;17:349–58.

[23] Wang B, Chen P, Huang DS, Li JJ, Lok TM, Lyu MR. Predicting protein interaction sites from residue spatial sequence profile and evolution rate. FEBS Lett 2006;580:380–4.

[24] Xia J-F, Zhao X-M, Song J, Huang D-S. APIS: accurate prediction of hot spots in protein interfaces by combining protrusion index with solvent accessibility. BMC Bioinf 2010;11:1.

[25] Li D, Ju Y, Zou Q. Protein folds prediction with hierarchical structured SVM. Curr Proteomics 2016;13:79–85.

[26] Zhu PP, Li WC, Zhong ZJ, Deng EZ, Ding H, Chen W, et al. Predicting the subcellular localization of mycobacterial proteins by incorporating the optimal tripeptides into the general form of pseudo amino acid composition. Mol Biosyst 2017;11:558–63, http://dx.doi.org/10.1039/c4mb00645c 2015.

[27] Wei L, Liao M, Gao X, Zou Q. Enhanced protein fold prediction method through a novel feature extraction technique. IEEE Trans Nanobiosci 2015;14:649–59.

[28] Zhao YW, Lai HY, Tang H, Chen W, Lin H. Prediction of phosphothreonine sites in human proteins by fusing different features. Sci Rep 2016;6:34817, http://dx.doi.org/10.1038/srep34817.

[29] Liu WX, Deng EZ, Chen W, Lin H. Identifying the subfamilies of voltage-gated potassium channels using feature selection technique. Int J Mol Sci 2014;15:12940–51, http://dx.doi.org/10.3390/ijms150712940.

[30] Lin H, Ding H, Guo FB, Zhang AY, Huang J. Predicting subcellular localization of mycobacterial proteins by using Chou's pseudo amino acid composition. Protein Pept Lett 2008;15:739–44.

[31] Liu B, Wang S, Dong Q, Li S, Liu X. Identification of DNA-binding proteins by combining auto-cross covariance transformation and ensemble learning. IEEE Trans Nanobiosci 2016;15:328–34.

[32] Liu B, Fang L, Liu F, Wang X, Chou K-C. iMiRNA-PseDPC: microRNA precursor identification with a pseudo distance-pair composition approach. J Biomol Struct Dyn 2016;34:220–32.

[33] Klein P. Prediction of protein structural class by discriminant analysis. Biochim Biophys Acta 1986;874:205–15.

[34] Huang D-S. A constructive approach for finding arbitrary roots of polynomials by neural networks. IEEE Trans Neural Netw 2004;15:477–91.

[35] Zhao Z-Q, Huang D-S, Sun B-Y. Human face recognition based on multi-features using neural networks committee. Pattern Recogn Lett 2004;25:1351–8.

[36] Huang D-S, Du J-X. A constructive hybrid structure optimization methodology for radial basis probabilistic neural networks. IEEE Trans Neural Netw 2008;19:2099–115.

[37] Ding YS, Zhang TL. Using Chou's pseudo amino acid composition to predict subcellular localization of apoptosis proteins: an approach with immune genetic algorithm-based ensemble classifier. Pattern Recogn Lett 2008;29:1887–92.

[38] Damoulas T, Girolami MA. Probabilistic multi-class multi-kernel learning: on protein fold recognition and remote homology detection. Bioinformatics 2008;24:1264–70, http://dx.doi.org/10.1093/bioinformatics/btn112.

[39] Dehzangi A, Phon-Amnuaisuk S, Dehzangi O. Using random forest for protein fold prediction problem: an empirical study. J. Inf. Sci. Eng 2010;26:1941–56.

[40] Zhao X, Zou Q, Liu B, Liu X. Exploratory predicting protein folding model with random forest and hybrid features. Curr Proteomics 2014;11:289–99.

[41] Zhang C-J, Tang H, Li W-C, Lin H, Chen W, Chou K-C. iOri-Human identify human origin of replication by incorporating dinucleotide physicochemical properties into pseudo nucleotide composition. Oncotarget 2017;7:69783–93, http://dx.doi.org/10.18632/oncotarget.11975 2016.

[42] Liu B, Long R, Chou K-C. iDHS-EL: identifying DNase I hypersensitive sites by fusing three different modes of pseudo nucleotide composition into an ensemble learning framework. Bioinformaitcs 2016;32:2411–8.

[43] Cao D-S, Zeng M-M, Yi L-Z, Wang B, Xu Q-S, Hu Q-N, et al. A novel kernel Fisher discriminant analysis constructing informative kernel by decision tree ensemble for metabolomics data analysis. Anal Chim Acta 2011;706:97–104.

[44] Wright J, Yang AY, Ganesh A, Sastry SS, Ma Y. Robust face recognition via sparse representation. IEEE Trans Pattern Anal Mach Intell 2009;31:210–27.

[45] Hu J, Han K, Li Y, Yang JY, Shen HB, Yu DJ. TargetCrys protein crystallization prediction by fusing multi-view features with two-layered SVM. Amino Acids 2016;48:2533–47.

[46] Jiang X, Lai J. Sparse and dense hybrid representation via dictionary decomposition for face recognition. IEEE Trans Pattern Anal Mach Intell 2015;37:1067–79.

[47] Xu Y, Zhang D, Yang J, Yang JY. A two-Phase test sample sparse representation method for use with face recognition. IEEE Trans. Circ. Syst. Video Technol 2011;21:1255–62.

[48] Xu Y, Zuo W, Fan Z. Supervised sparse representation method with a heuristic strategy and face recognition experiments. Neurocomputing 2012;79:125–31.

[49] Yang J, Chu D, Zhang L, Xu Y, Yang J. Sparse representation classifier steered discriminative projection with applications to face recognition. IEEE Trans Neural Netw Learn Syst 2013;24:1023–35.

[50] Huang K, Aviyente S. Advances in Neural Information Processing Systems; 2017. p. 609–16.

[51] Huang Y-A, You Z-H, Gao X, Wong L, Wang L. Using weighted sparse representation model combined with discrete cosine transformation to predict protein-protein interactions from protein sequence. BioMed Res Int 2015;2015.

[52] Yu D-J, Hu J, Huang Y, Shen H-B, Qi Y, Tang Z-M, et al. TargetATPsite: a template- free method for ATP-binding sites prediction with residue

evolution image sparse representation and classifier ensemble. J Comput Chem 34 2013:974–85.

[53] Taguchi Y, Gromiha MM. Application of amino acid occurrence for discriminating different folding types of globular proteins. BMC Bioinf 2007;8:1.

[54] Lyons J, Dehzangi A, Heffernan R, Yang YD, Zhou YQ, Sharma A, et al. Advancing the accuracy of protein fold recognition by utilizing profiles from hidden Markov models. IEEE Trans Nanobiosci 14 2015:761–72.

[55] Sharma A, Lyons J, Dehzangi A, Paliwal KK. A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition. J Theor Biol 2013;320:41–6.

[56] Shen H-B, Chou K-C. Predicting protein fold pattern with functional domain and sequential evolution information. J Theor Biol 2009;256:441–6.

[57] Donoho DL. Compressed sensing. IEEE Trans Inf Theory 2006;52:1289–306.

[58] Candès EJ, Romberg J, Tao T. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. IEEE Trans Inf Theory 2006;52:489–509.

[59] Chang C-C, Lin C-J. LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol (TIST) 2011;2:27.

[60] John GH, Langley P. Proceedings of the Eleventh conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc.; 2017. p. 338–45.

[61] Shamim MTA, Anwaruddin M, Nagarajaram HA. Support Vector Machine-based classification of protein folds using the structural properties of amino acid residues and amino acid residue pairs. Bioinformatics 2007;23:3320–7.

[62] Saigo H, Vert J-P, Ueda N, Akutsu T. Protein homology detection using string alignment kernels. Bioinformatics 2004;20:1682–9.

[63] Rangwala H, Karypis G. Profile-based direct kernels for remote homology detection and fold recognition. Bioinformatics 2005;21:4239–47.

[64] Yang JY, Chen X. Improving taxonomy-based protein fold recognition by using global and local features. Proteins: Struct Funct Bioinf 2011;79:2053–64.

[65] Lin C, Zou Y, Qin J, Liu XR, Jiang Y, Ch. Ke, et al. Hierarchical classification of protein folds using a novel ensemble classifier. PLoS One 2013;8, e56499.

[66] Chen D, Tian X, Zhou B, Gao J. Profold: protein fold classification with additional structural features and a novel ensemble classifier. BioMed Res Int 2016;2016.

[67] Ghanty P, Pal NR. Prediction of protein folds: extraction of new features, dimensionality reduction, and fusion of heterogeneous classifiers. IEEE Trans Nanobiosci 2009;8:80–110.

[68] Sharma A, Paliwal KK, Dehzangi A, Lyons J, Imoto S, Miyano S. A strategy to select suitable physicochemical attributes of amino acids for protein fold recognition. BMC Bioinf 2013;14(1).

[69] Paliwal KK, Sharma A, Lyons J, Dehzangi A. A tri-gram based feature extraction technique using linear probabilities of position specific scoring matrix for protein fold recognition. IEEE Trans Nanobiosci 2014;13:44–50.

[70] Lyons J, Biswas N, Sharma A, Dehzangi A, Paliwal KK. Protein fold recognition by alignment of amino acid residues using kernelized dynamic time warping. J Theor Biol 2014;354:137–45.

[71] Paliwal KK, Sharma A, Lyons J, Dehzangi A. Improving protein fold recognition using the amalgamation of evolutionary-based and structural based information. BMC Bioinf 2014;15:S12.